PROJECT 1: LINEAR REGRESSION
MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING), 2018
Peer assessment: Monday 23 April (13–14)
Final deadline: Thursday 26 April, 16.00

1 Instructions

You must write a report, in English. Work in groups of two. Discussion between groups is permitted (and encouraged), as long as your report reflects your own work. Write a clear report, presenting your approach to the assignment, discussing methods and results. Results discussion and interpretation is important. Just reporting results is not enough! It should be noted that for some questions there isn’t a unique "right" answer and there are a myriad of different issues that you could discuss, so use your imagination. In addition to the text, use as many figures and tables as necessary, with explanatory captions.

The report should be readable, not a random disorganized collection of thoughts, plots and tables (see also the Peer Review Guidelines at the end of this document). For example, it should be possible for the reader to understand what you are doing without having access to your code. Also, key information may be better summarized in tables than by including the R printouts (e.g. it may be enough to give regression coefficients and p-values without all the accompanying information provided by R). There is no need to include your R code in the report, but you can include some of the R output.

1.1 Form groups

Form groups of two students via “SAM”: https://matstat.sam.cs.lth.se/Labs

1.2 Peer review

Bring a printed version of your report to the peer assessment (23 April, 13), or email the report to anna@maths.lth.se at least 1 hr in advance so I can print a copy.

1.3 Final submission

E-mail the final version (a single PDF document) to one of (depending on your course) the following addresses by the deadline 16.00 at Thursday 26 April. Also attach to the same message your R-files (or implementation in other language), in a file named proj1.R that can be used to run your analyses.

- MASM22/FMSN30 students: email to fmsn30@matstat.lu.se
- FMSN40 students: email to fmsn40@matstat.lu.se

Subject field of the email: write “Project1 by studid1 and studid2” where studid1 and studid2 are the id numbers for two students in a given group (forgot your id? Go to the link in the “Form groups” section).

Example: Project1 by d08xhj and d08fjh
2 Determinants of Plasma Retinol and Beta-Carotene Levels

2.1 Summary

Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. We designed a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. Study subjects \((N = 315)\) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. We display the data for only two of the analytes.

We conclude that there is wide variability in plasma concentrations of these micronutrients in humans, and that much of this variability is associated with dietary habits and personal characteristics. A better understanding of the physiological relationship between some personal characteristics and plasma concentrations of these micronutrients will require further study.

2.2 Reference:


2.3 Description:

The datafile contains 315 observations on 14 variables available and is taken from http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/plasma.html

The data file plasma.txt can be downloaded from the course home page. Save it to your R working directory and then read it into R and put it in a data frame called plasma with

```r
plasma <- read.delim("plasma.txt")
```

2.4 Variables:

- **age** Age (years)
- **sex** Sex (1=Male, 2=Female)
- **smokstat** Smoking status (1=Never, 2=Former, 3=Current Smoker)
- **quetelet** Quetelet (weight/(height^2))
- **vituse** Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
- **calories** Number of calories consumed per day.
- **fat** Grams of fat consumed per day.
- **fiber** Grams of fiber consumed per day.
- **alcohol** Number of alcoholic drinks consumed per week.
- **cholesterol** Cholesterol consumed (mg per day).
- **betadiet** Dietary beta-carotene consumed (mcg per day).
- **retdiet** Dietary retinol consumed (mcg per day).
- **betaplasma** Plasma beta-carotene (ng/ml)
- **retplasma** Plasma Retinol (ng/ml)
3 Questions

3.1 Plasma Retinol and Age

(a). We want to model how Plasma Retinol varies with Age. Plot them against each other and determine, visually, whether a linear relationship might be appropriate.

(b). Fit the model \( \text{retplasma} = \beta_0 + \beta_1 \cdot \text{age} + \varepsilon \) and report the \( \beta \)-estimates together with their 95% confidence intervals. Are the parameters significant?

(c). What happens, on average, to the Plasma Retinal level if we increase the age by 1 year? Calculate a 95% confidence interval for the change. Does the size of the change depend on the age?

(d). Investigate the residuals. Do they fulfill the model assumptions? If not, what seems to be the problem?

(e). Ignoring any problems with the residuals, calculate the 95% confidence interval for the expected Plasma Retinol for ages 18, . . . , 85 and add it to the plot of the data, together with the estimated relationship. Do the problems with the residuals seem to have an impact?

(f). Calculate the 95% prediction interval for the expected Plasma Retinol for ages 18, . . . , 85 and add it to the plot. Do you find any problems here? How does this relate to the problem with the residuals?

(g). Report a 95% prediction interval for the observed Plasma Retinol of a 30 year old person, as well as for a 70 year old person. Are there any substantial differences in the widths of the two intervals? Why or why not?

3.2 log Plasma Retinol and age

(a). Investigate whether taking the logarithm of the Plasma Retinol might improve the model fit and reduce the problems with the residuals by fitting the new model \( \log(\text{retplasma}) = \beta_0 + \beta_1 \cdot \text{age} + \varepsilon \). Report the estimates and their confidence intervals. Are they significant?

(b). Investigate the new residuals. Do they fulfill the model assumptions? Did the transformation solve the problems?

(c). Write down how Plasma retinol depends on age and plot the data again, adding this new, non-linear, fit. What happens, on average, to the Plasma Retinal level if we increase the age by 1 year? Calculate a 95% confidence interval for this change. Does the size of the change (in ng/ml) depend on the age?

(d). Calculate the 95% confidence interval for the expected log Plasma Retinol for ages 18, . . . , 85 and transform it into the original scale. Add it to the plot of the data, Any major differences compared to the previous model? Why or why not?

(e). Calculate the 95% prediction interval for the expected log Plasma Retinol for ages 18, . . . , 85, transform it into the original scale and add it to the plot. Do you find any problems here now? Any major differences compared to the previous model?

(f). Report a 95% prediction interval for the observed Plasma Retinol of a 30 year old person, as well as for a 70 year old person. Are there any substantial differences in the widths of the two intervals? Why or why not?
3.3 Plasma Beta-carotene — checking for possible problems

We now move from Plasma retinol to Plasma beta-carotene. We want to model how (a suitable transformation of) Plasma Beta-carotene varies as a function of the background variables (age, sex, smokstat and quetelet = BMI) and the dietary factors (vituse, calories, fat, fiber, alcohol, cholesterol and betadiet).

We start by examining the relationships with some of the explanatory variables, checking for possible problems.

(a). Start by plotting and fitting a simple linear model with only age as explanatory variable. Examine the residuals and decide whether you should transform the Plasma beta-carotene in some way. Then use that transformation in all of the other analyses below.

*Note*: This might cause a problem with one observation. Which one and why? In order of do the analyses without this problematic observation you can create a new data frame using

\[
\text{newdata} \leftarrow \text{subset(olddata, variable>0)}
\]

This will create a new data frame using only the rows where variable is greater than 0.

(b). Fit a model using only the Smoke status variable as explanatory variable(s), with "Never" as reference category. Note the parameter estimates and their standard errors.

Then fit a model using "Current smoker" as reference category instead. Note the parameter estimates and their standard errors. What happened to the standard errors? Explain why.

Now check all the other categorical variables to find out whether the first category is a suitable reference category. Does it matter which sex category we pick as reference? Why/why not?

(c). Plot (the transformed) Plasma beta-carotene against Alcohol. Do you see any potential problems here? Would it be a good idea to take the logarithm of the alcohol consumption? Why not? *Hint*: how many persons do not consume any alcohol at all?

Also plot the other explanatory variables against alcohol. Is the extreme person (drinking the equivalent of 1 liter of vodka a day!) extreme in any other variables as well?

*Note*: We will use the alcohol variable as it is in the models but we will keep track of the extreme value and see what problems it causes later, if any.

(d). Plot all the continuous explanatory variables against each other. Are there any variable pairs that look like they might cause colinearity problems?

3.4 Plasma Beta-carotene — multiple regression

We are now ready to start building models. Keep in mind all the potential problems, and their solutions, you identified in 3.3. Questions marked with * requires material from Lectures 5 and 6.

(a). Fit a model using the background variables (age, sex, smokstat and quetelet), report the parameter estimates and their corresponding 95 % confidence intervals. Also determine whether all the variables have a significant contribution to the model.

Use the fitted model to construct an interval that would be expected to contain 95 % of the Plasma beta-carotene values of persons that are 50 years old, male, have never smoked and have a BMI of 30.

(b). * Fit a model using the dietary factors (vituse, calories, fat, fiber, alcohol, cholesterol and betadiet) instead. Are all the variables significant? If not, use a stepwise procedure to reduce the model. Report the parameter estimates and the corresponding confidence intervals of the reduced model.
Use the reduced model to construct an interval that would be expected to contain 95% of the Plasma beta-carotene values of persons that never eat vitamins, consume 1200 calories, 50 grams of fat, 20 grams of fiber, no alcohol, 300 mg cholesterol and 1500 mcg dietary beta-carotene per day.

(c). We now have two competing models, the background variables model in [a] and the (reduced) dietary factors model in [b]. Compare the two models regarding, e.g., their ability to explain the variability in Plasma beta-carotene. Which model seems best?

Try to find a better model using both some of the background variables and some of the dietary factors. Compare its ability to explain the variability to the models from [a] and [b].

(d). Now we turn our attention to the, possibly, problematic extreme alcohol consumer by investigating the leverage, studentized residuals, Cook's distance and DFbetas of the (reduced) dietary factors model in [b]. Has the person had any problematic influence on the model estimates? Are there any other persons that have had a problematic influence?

End of Project 1
4 Peer review guidelines

The following guidelines for peer-review should be followed.

4.1 Questions regarding the content

1. Have all the tasks in the assignment been completed? □ □
2. Does the report contain relevant figures and tables? □ □
3. Has all notation been properly introduced and/or explained? □ □
4. Has the model been properly introduced? □ □
5. Are the results properly presented and discussed? □ □

4.2 Questions regarding the report presentation

1. Does the report have:
   • Title, authors, and date? □ □
   • Page numbers? □ □
   • Introduction? □ □
   • Results and/or conclusions? □ □
2. Has the report been proofread? Have language and spelling mistakes been corrected? □ □
3. Are figures and tables:
   • Numbered? □ □
   • Equipped with suitable captions? □ □
   • Referred to in the text? □ □
4. Is the text divided into paragraphs and well structured with clear and suitable section headings? □ □
5. Is the report easy to read, and understandable without access to the project description? □ □